PROCESSING OF SEMANTIC INFORMATION IN FLUENTLY SPOKEN LANGUAGE

Allen L. Gorin

AT&T Research
Murray Hill, New Jersey
algor@research.att.com

ABSTRACT

We are interested in constructing machines which learn to understand and act upon fluently spoken input. For any particular task, certain linguistic events are critical to recognize correctly, others not so. This notion can be quantified via *salience*, which measures the information content of an event for a task. In previous papers, salient words have been exploited to learn the mapping from spoken input to machine action for several tasks. In this work, a new algorithm is presented which *automatically acquires salient grammar fragments* for a task, exploiting both linguistic and extra-linguistic information in the inference process. Experimental results will be reported for a database of fluently spoken customer requests to operators, responding to the open-ended prompt of '*Hello, this is AT&T. How may I help you?*'

INTRODUCTION

There is a large body of research in spoken language understanding systems, almost all of which are predicated upon the user saying what the device expects to hear. In those systems, it is the user's responsibility to learn the machine's vocabulary and grammar in order to achieve reasonable performance. In contrast, this research program focuses on shifting that burden from human to machine, making it the device's responsibility to respond appropriately to what people actually say.

One approach to extracting semantic information from fluent natural language is to search for meaningful fragments and combinations thereof. Such systems are based on the intuition that some linguistic events are critical to recognize for a particular task, others not so. We have quantified this notion via *salience*, which measures the information content of an event for a task [G95]. In previous work, we have exploited salient words to understand and act upon unconstrained input for a variety of tasks [G94a] [H94][M93][S93]. In this new work, we describe an inference algorithm which *automatically acquires salient grammar fragments* from a database of transcribed utterances labeled with associated machine actions. While there is a large literature on self-organizing language models, such efforts have traditionally exploited only the language itself, with the goal of within-language prediction to improve ASR.

This is actually a much harder problem than people are faced with, who acquire language during the course of interacting with a complex environment. This new algorithm, following that intuition, exploits both language and extra-linguistic information to infer structure.

COMMUNICATION AND SALIENCE

Consider devices whose purpose is to understand and act upon fluently spoken input. The goal of communication in such systems is to induce the machine to perform some action or to undergo some internal transformation. The communication is judged to be successful if the machine responds appropriately. We have explored this paradigm in some detail [G95], in particular contrasting it with traditional communication theory, where the goal is to reproduce a signal at some distant point.

We have constructed several devices which acquire language via building associations between input stimuli and appropriate machine responses tasks [G93][G94a][M93][S93][H94]. The *meaning* of an input stimulus (e.g. a word) can be defined [G95] via its statistical associations to a device's input/output periphery. This has the attractive property of grounding meaning in a device's experiences and interactions with its environment. Viewing this set of associations as a vector enables one to define a *semantic distortion* between events as the distance between their association vectors. The salience of an event is then defined as its distance from a null-event.

In the case that associations are defined via mutual information between events, then this semantic distortion can be shown to be equivalent to the relative entropy between the *a posteriori* distributions of the output actions (conditioned upon the two input events). The salience of an event is then the unique non-negative measure of how much information that event provides about the random variable of appropriate machine responses. The reader is referred to the tutorial paper in [G95] for a detailed discussion of these ideas.

TASK AND DATABASE

In previous work, we have described the task of learning to understand what customers say to telephone operators, responding to the prompt of *'Hello, this is AT&T. How may I help you?'* [G94b]. The first stage in such a system must determine which of several services a user desires. The call can then be routed to either an automated subsystem or an appropriate human agent. For

example, if the caller says '*I want to reverse the charges on this call*', then an appropriate response would be to connect them to the existing subsystem for automating collect calls. Another example might be '*I can't understand my phone bill*', whence the caller should be transferred to the business office.

As detailed in [G96], we have created a database of more than 10K such transactions, based on approximately 200 hours of recorded dialogs between customers and operators. While the language usage is highly variable (comprising nearly 4K words), most people are asking for one of 15 services. Some are information queries, such as '*What's the area code of Chicago*?'. Others are service requests, such for a billing credit, as in '*I dialed the wrong number*.' The experiments reported in this paper are based on an early subset of that database, comprising the first customer utterance, an orthographic transcription thereof, and a call-type label indicating the service requested.

SALIENT PHRASE FRAGMENTS

In previous work, we introduced the notion of a salient word, demonstrating that a rudimentary understanding module can be constructed based on only that subset. For example, a salience analysis of the operator services corpus yields the following list

WORD	SALIENCE
difference	4.04
cost	3.39
rate	3.37
much	3.24
emergency	2.23
misdialed	1.43
wrong	1.37
code	1.36
dialed	1.29
area	1.28
time	1.23
person	1.23
charge	1.22
home	1.13
information	1.11
credit	1.11

Table 1: Some Salient Words

We now search the space of phrase fragments, guided by two criteria. First, within the language channel, a word pair v_1v_2 is considered as a candidate unit if it has high mutual information,

$$I(v_1,v_2) = \log_2 [P(v_2|v_1)/P(v_2)]$$

This measure can be composed to recursively construct longer units, by computing I(f,v) where f is a word-pair or larger fragment. We then introduce an additional between-channel criterion, which is that a fragment should have high information content for the call-action channel. Following [G95], where f is a fragment and $\{c_k\}$ is the set of call-actions, denote its salience by

$$S(f) = \Sigma\, P(c_k|f)I(f,c_k)$$

We perform a breadth-first search on the set of phrases, up to length four (an arbitrary cut-off point), pruning it by these two criteria: one defined wholly within the language channel, the other defined via the fragment's extra-linguistic associations. The following table illustrates some salient and background phrase fragments generated by this algorithm. Three attributes of each fragment are provided. First, the mutual information between the final word in the fragment and the preceding subfragment, denoted *MI*. Second, the peak of the *a posteriori* distribution $P(c_k|f)$, denoted P_{max}. Third, the call-type for which that peak occurs, denoted *Call-Type*. When the peak is between 0.5 and 0.9, then the fragment is only moderately indicative of that call-type and so is provided within parentheses. When the peak is low (<0.5), then it is a background fragment not associated with any particular call-type, so none is provided

MI	Phrase Fragments	P_{max}	Call-Type
7.4	made a long distance	0.93	Billing Credit
7.3	long distance	0.55	(Billing Credit)
7.1	I would like	0.24	
6.9	area code	0.65	(Area Code)
6.3	could you tell me	0.37	
5.6	the area code for	0.92	Area Code
5.3	I'm trying	0.33	
5.0	a wrong number	0.98	Billing Credit
4.9	a long distance call	0.62	(Billing Credit)
4.8	the wrong number	0.98	Billing Credit
4.8	a number in	0.29	
4.4	I'm trying to	0.33	
4.3	long distance call	0.62	(Billing Credit)
4.3	I just made a	0.93	Billing Credit
4.1	I'd like to	0.18	

Table 2: Salient and Background Phrase Fragments

For example, consider the fragment '*long distance*', which has a strong co-occurrence pattern within the language channel, thus a high mutual information (MI=7.3). However, it is not a very meaningful phrase in the sense that the most likely call-type (given that phrase in an utterance) is a billing credit query, but only with probability 0.55. Consider on the other hand an extension of that phrase, '*made a long distance*', which both has high mutual information (MI=7.4) and strongly connotes a billing credit query with probability 0.93. A similar discussion can be made for the fragments '*area code*' and '*the area code for*'. There are several background fragments in the list, which have strong co-occurrence

patterns but are not indicative of any particular call-type, such as '*I would like*' and '*could you tell me*'. Such fragments can still be useful for creating improved background models for speech recognition, which will be addressed in later research.

SALIENT GRAMMAR FRAGMENTS

We now consider a method for combining salient phrase fragments into a grammar fragment. For example, in Table 2, consider the two salient phrases '*a wrong number*' and '*the wrong number*'. Clearly, these should not be treated independently, but rather combined into a single unit. The key idea is that there are two similarity measures, one in the language channel, the other extra-linguistic. Within-channel, there are various measures to compute similarity of word-strings (e.g. a Levenshtein distance). We impose the extra-linguistic constraint, however, that in order for two strings to be merged, then their meaning must be similar.

Table 3 below illustrates the growth of a salient grammar fragment for billing credit queries. For simplicity of exposition, we restrict attention to a single call-type, focusing on distinguished billing credit queries from all others. The first pass of the algorithm determines salient words, for which the top choice (for billing credits) is '*wrong*'. The others are '*dialed*', '*credit*', '*disconnected*', '*misdialed*' and '*cut*'.

Prob corr **P(Cr | G)**	**Coverage** **P(G| Cr)**	**Fragment** **G**
0.92	0.48	wrong
0.98	0.41	wrong number
0.95	0.45	wrong (number|eos|call)
0.97	0.42	(a|the|was) wrong (number|eos|call)
0.95	0.50	F(wrong) | F(dialed)
0.95	0.57	F(wrong) | F(dialed) | F(credit)
0.95	0.59	--- | F(disconnected)
0.95	0.64	--- | F(misdialed) | F(cut off)

Table 3: Growth of a Salient Grammar Fragment for Distinguishing Billing Credit Queries

The word '*wrong*' is strongly indicative of billing credit (denoted *Cr*), with P(*Cr*|*wrong*)=0.92. The coverage is low, however, with only 48% of those queries containing that word. The local context of this salient word is then evaluated for those elements which sharpen the semantics, i.e. increase the classification rate. The top choice for expanding local context is then '*wrong number*', which sharpens the *a posteriori* probability to 0.98. Similarly, other left and right contexts are added, leading to the grammar fragment

F(wrong) = (a|the|was) wrong (number|eos|call),

where *eos* is the end-of-sentence marker, | indicates disjunction (or) and concatenation indicates conjunction in order. The fragment with the kernel '*wrong*' is then denoted F(*wrong*). At this point, the semantics is quite sharp, with the *a posteriori* probability being 0.97, although the coverage has dropped to 0.42. This process is then repeated to construct fragments surrounding the other salient words for this call-type, denoted F(dialed), etc. As this expression becomes too long to fit in the table, we indicate the fragment from the previous row by '---'. By incrementally adding these fragments, the coverage is increased to 0.64 while maintaining a high classification rate of 0.95.

EXPERIMENTAL RESULTS

Consider the two-class problem of distinguishing billing credit queries from the others. For any particular salience threshold, a particular set of grammar fragments will be generated. A most rudimentary decision rule would be based simply whether one of these fragments matches a substring of the recognizer output.

For example, the following are some illustrative correct detections of a billing credit query, based on such a matching scheme. The substring which matches a grammar fragment is highlighted by capitalization plus connection with underscores. Digit sequences are denoted '*xxx*'.

Correct Detections

- i placed a call and i GOT_A_WRONG_NUMBER earlier this afternoon.
- yes i MISDIALED a number.
- I_WAS_CUT_OFF when trying to call this number.
- I_WAS_DIALING 1 xxx xxx xxxx and i got someone else
- yes I_JUST_DIALED AN_INCORRECT_NUMBER
- yes I would like TO_GET_CREDIT_FOR a number I called

There are two types of errors that occur in such a classifier. First is a *false detection*, i.e. classifying a call as a billing credit when it was not. Second is a *missed detection*, i.e. a billing credit query that was classified as other. The operational costs of such errors can be quite different. For example, a missed detection in a call-router leads to a missed opportunity for automation, while a false detection leads to an incorrect routing. Several examples of such errors are shown below.

False Detections

- yes i have a number here and i don't know if it's A_WRONG_NUMBER

- I was trying to get xxx xxx xxxx and it said it WAS_DISCONNECTED

Missed Detections

- I am trying to call wooster and the number I have rings to a different number
- I'm going to blame this one on my wife I misread her handwriting
- I'm dialing xxx xxx xxxx and I keep getting bells and things like that

An experiment was conducted on a small subset of the speech corpus comprising 1800 utterances. A statistical bigram language model was trained on a separate corpus of 2200 transcriptions, which were also used to train salient grammar fragments for billing credit queries. An HMM-based large-vocabulary recognizer (BLASR) was used with off-the-shelf subword models and standard dictionary pronunciations for the vocabulary present in the training set (~2000 words). Call-type classification was implemented via spotting for salient fragments in the recognizer output. Performance was varied by selecting different salience thresholds, trading probability of correct rejection against correct detection. The performance curve for spoken input is shown in Figure 1 below. For comparison, the performance of the billing credit detector is also shown on transcribed (text) input.

CONCLUSIONS

In conclusion, we have described a new algorithm for automatically acquiring salient grammar fragments from a corpus, exploiting both linguistic and extra-linguistic information. Preliminary performance results have been reported for distinguishing billing credit queries from others from a database of fluently spoken customer responses to the prompt '*Hello, this is AT&T. How may I help you?*

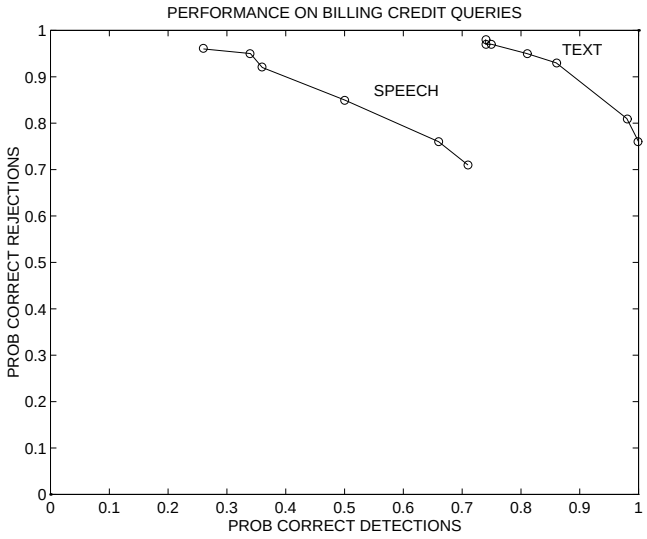

Figure 1: Distinguishing Billing Credit Queries versus Other Call-Types from Spoken Input

REFERENCES

[G93] A.N. Gertner and A.L. Gorin, "*Adaptive Language Acquisition for an Airline Information Subsystem*," Artificial Neural Networks for Speech and Vision, pp. 401-428, (ed. R. Mammone), Chapman and Hall, 1993.

[G94a] A.L. Gorin, S.E. Levinson, and A. Sankar "*An Experiment in Spoken Language Acquisition*," IEEE Trans. on Speech and Audio, pp. 224-240, vol. 2, no. 1, part II, Jan. 1994.

[G94b] A.L. Gorin, H. Hanek, R. Rose and L. Miller," *Spoken Language Acquisition for Automated Call Routing*," Proc. of the Intl. Conf. on Spoken Language Processing, pp. 1483-1485, Sept. 1994, Yokohama, Japan.

[G95] A.L. Gorin, "*On Automated Language Acquisition*," 97(6), pp. 3441-3461, Journal of the Acoustical Society of America (June 1995).

[H94] E.A. Henis, S.E. Levinson and A.L. Gorin, "*Mapping Natural Language and Sensory Information into Manipulatory Actions*," Proc. of the eighth Yale workshop on adaptive and learning systems, June 1994.

[M93] L.G. Miller and A.L. Gorin, "*Structured Networks for Adaptive Language Acquisition*," International Journal of Pattern Recognition and Artificial Intelligence, special issue on Neural Networks, vol. 7, no. 4, pp. 873-898 (1993).

[S93] A. Sankar and A.L. Gorin, "*Visual Focus of Attention in Adaptive Language Acquisition*," Artificial Neural Networks for Speech and Vision, pp. 324-356, (ed. R. Mammone), Chapman and Hall, 1993.

[G96] A.L. Gorin, B.A. Parker, R.M. Sachs and J.G. Wilpon, '*How may I help you*?', to appear in the Proc. of IVTTA 1996.